Psycholinguistic Experiments Measuring the Effects of MDE on Readability

Rich Transcription RT03-F Workshop

13 November 2003 16:05-16:35

Ted Gibson, Florian Wolf, Evelina Fedorenko and Doug Jones

egibson@mit.edu, fwolf@mit.edu, evelina9@mit.edu, daj@ll.mit.edu

Brain and Cognitive Science Department and Lincoln Laboratory

This work was sponsored by the Defense Advanced Research Projects Agency (DARPA) under Air Force Contract number F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

Outline for Readability Update

- Passage-level readability studies
  - 2x2 design for each experiment: (system, reference) x (clean-up, no clean-up)
  - Experiment 1: Passage + questions presented at once
  - Experiment 2: Passage, then questions presented later
  - NEW – Experiment 3: Same as 2, without speaker information
- NEW Moving-window readability studies
  - Experiment 4: Three-word window
  - Experiment 5: One-word window
- General results: Cleanup improves readability

Measuring Readability

(1) **Reading Time**

(2) **Answers to Comprehension Questions**

(3) **Subjective Difficulty Rating**

(4) **(Question Answering Time)**

How many days per week does one of the speakers exercise?

(a) one
(b) three
(c) five
(d) every second day
(e) every day

Eurospeech 2003

Three Stages of STT Enhancement

STT-	STT+	XT
w d w d w w w d w w w w d w w w d w w w	0: w d w d 1: w w w d 0: w w w 1: w d w w	0: W w? 1: W w w. 0: W w w. 1: W w w?
Temporally ordered words (includes disfluencies).	Add diarization / speaker segmentation.	Remove disfluencies, add punctuation/capitalization.

Text Conditions for Passage-Level Designs

STT+(ref)

1: yeah actually um i belong to a gym down here a gold's gym and uh exercise i try to exercise five days a week um and i usually do that uh […]

0: what type of exercising do you do in the gym

XT(ref)

1: Yeah I belong to a gym down here. Gold's Gym. And I try to exercise five days a week. And I usually do that. […]

0: What type of exercising do you do in the gym?

STT+(sys)

1: actually uh i belong to a gym down here a gold jim uh i exercise so i tried exercise five days a week uh i usually do that […]

0: took said can you imagine

XT(sys)

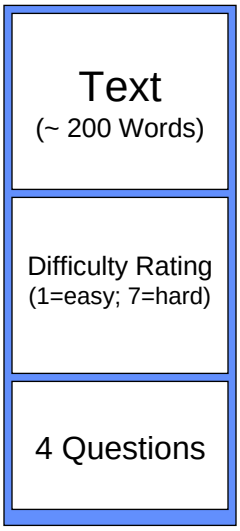
1: I belong to a gym down here a gold jim. And I exercise so I tried exercise five days a week. And I usually do that. […]

0: What took said can you imagine.

Eurospeech 2003

Passage-level Forms for Test Trials

- **All Trials**
  - **Subject is presented with text and questions**
  - **Reading is timed**
  - **Subject rates reading difficulty for text**
  - **Answers to comprehension questions are recorded**
- **Form I: text and questions appear together**
  - **Simulates scanning task**
  - **Much easier**
- **Form II: text appears first, then disappears, then questions appear.**
  - **General reading task**
  - **Significantly harder**



Eurospeech 2003

Passage-level Experiments

- **Results:**
  - **Reference texts are processed better than system texts (3 measures: Response accuracy to questions; Subjective text difficulty ratings; and Reading Times (RTs))**
  - **Cleaned-up reference texts are rated as better than raw reference texts; no differences in more objective measures (RTs, Response accuracies)**
- **Presented at EARS May 2003 PI meeting & Eurospeech 2003**

Outline for Readability Update

- Passage-level readability studies
  - 2x2 design for each experiment: (system, reference) x (clean-up, no clean-up)
  - Experiment 1: Passage + questions presented at once
  - Experiment 2: Passage, then questions presented later
  - NEW – **Experiment 3: Same as 2, without speaker information**
- NEW **Moving-window readability studies**
  - **Experiment 4: Three-word window**
  - **Experiment 5: One-word window**
- **General results: Cleanup improves readability**

Experiment 3: No speaker information

- Same 2x2 materials:
  (raw=STT-, cleaned-up=XT) x (reference, system)
- No speaker-change information

- Results: Same as Experiments 1 and 2:
  - Reference texts are processed better than system texts
  - Cleaned-up reference texts are rated as better than raw reference texts; no differences in more objective measures

- Importantly: No difference between this experiment and Experiments 1 and 2 on any measure (reading times, response accuracy, difficulty ratings)
  - People responded to questions just as accurately without speaker-change information as with speaker-change information.
  - Ratings were the same across the experiments.

Experiment 3: No speaker information

- Possible reasons for no differences with Experiments 1 and 2:
  - Ceiling effects in response accuracies, reading times.
  - Ratings are always relative to other materials in an experiment:
    +/- speaker-change information was not manipulated in this experiment

- Follow-up: Test speaker-change information within an experiment, not across experiments, using more fine-grained methods.
  - Such an experiment is currently running.

Outline for Readability Update

- Passage-level readability studies
  - 2x2 design for each experiment: (system, reference) x (clean-up, no clean-up)
  - Experiment 1: Passage + questions presented at once
  - Experiment 2: Passage, then questions presented later
  - NEW – Experiment 3: Same as 2, without speaker information

- NEW **Moving-window readability studies**
  - **Experiment 4: Three-word window**
  - **Experiment 5: One-word window**

- **General results: Cleanup improves readability**

Moving-window readability studies

- **Experiment 4: self-paced reading, 3 words at a time.**
- **New materials: 18 texts from 40 Meteer-annotated conversations from Penn Treebank-3 (courtesy of Liz Shriberg)**
- **New mark-up of texts, with two goals in mind (in progress):**
  - **Testing readability of transcripts;**
  - **Ease and cross-coder reliability of coding texts.**
- **Three categories of mark-up (cf. MDE5):**
  - **Punctuation: periods, capitalization, commas, question marks.**
  - **Filler: filled pauses (uh, um, etc.), verbatim repeated words and phrases, discourse markers ("you know", "anyways", "like")**
  - **Edits: Omit material that makes the conversation incoherent. E.g., re-starts, complex edits.**
- **Mark-up in current experiments: Consensus among 3 annotators**

Conditions in Experiment 4

- **6 conditions: All reference texts (no system texts)**

1. **Raw text**
2. **Raw text + punctuation only**
3. **Raw text, no fillers**
4. **Raw text, no fillers, no edits**
5. **Raw text + punctuation, no fillers, no edits, but leaving in empty speaker turns (lost in filtering)**
6. **Fully cleaned up text: Raw text + punctuation, no fillers, no edits, no empty speaker turns**

Experiment 4 and 5: Materials

```
A: it's in dublin texas i'm not really sure
what the county it's stephen no
stephenville
B: oh okay okay
A: okay where stephenville's at
B: uhhuh
A: they've had a lot of problems because
they've um introduced a lot of dairies
there
B: yeah
```

Experiment 4: Self-paced, 3 words at a time

A: ---- -- ------ ----- --- --- ----- ----
---- --- ----- --- ------- -- ----------
B: -- ---- ----
A: ---- ----- ---------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time

A: it's in dublin ----- --- --- ----- ----
---- --- ----- --- ------- -- ----------
B: -- ---- ----
A: ---- ----- ---------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time

A: ---- in dublin texas --- --- ----- ----
---- --- ----- --- ------- -- -------------
B: -- ---- ----
A: ---- ----- ------------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time

A: ---- -- dublin texas i'm --- ----- ----
---- --- ----- --- ------- -- -------------
B: -- ---- ----
A: ---- ----- ------------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time

A: ---- -- ------ texas i'm not ----- ----
---- --- ----- --- ------- -- -------------
B: -- ---- ----
A: ---- ----- ------------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

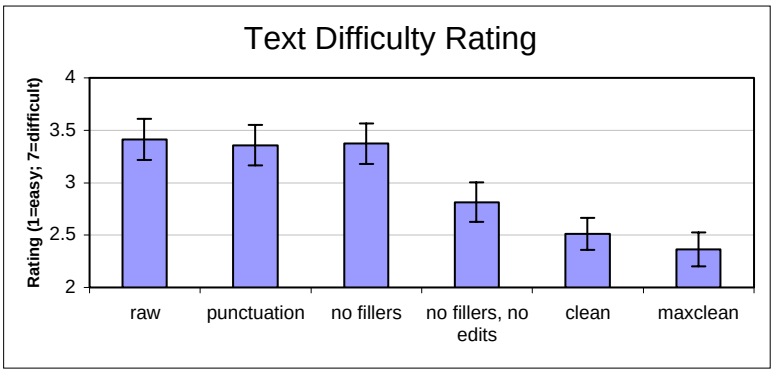
Experiment 4: Self-paced, 3 words at a time

A: ---- -- ------ ----- i'm not really ----
---- --- ----- --- ------- -- -------------
B: -- ---- ----
A: ---- ----- ------------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time

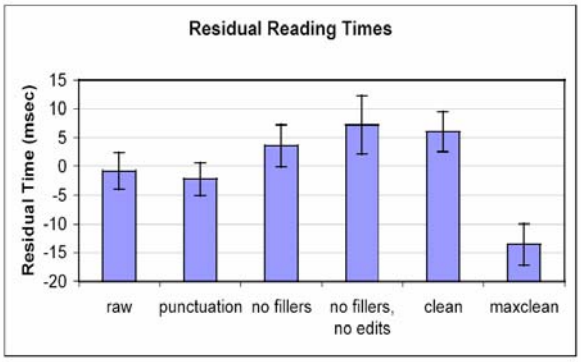
A: ---- -- ------ ----- --- not really sure
---- --- ----- --- ------- -- -------------
B: -- ---- ----
A: ---- ----- ------------- --
B: -----
A: ------- --- - --- -- -------- ------- --
----- -- ---------- - --- -- ------- -----
B: ----
...

Experiment 4: Self-paced, 3 words at a time



- **Including punctuation or taking away fillers by themselves do not improve readability.**
- **Taking away fillers+edits improves readability.**
- **Putting in punctuation improves readability once the edits are removed.**

Experiment 4: Self-paced, 3 words at a time



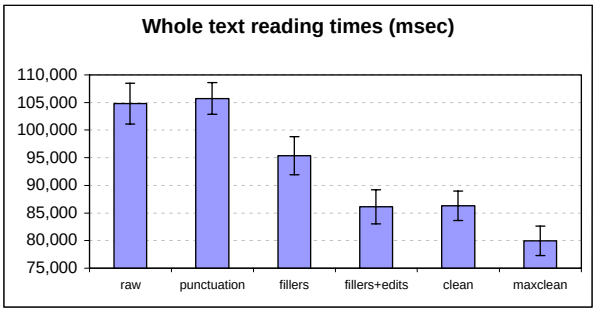
Result 1: Cleaned-up texts are read ~15 msec/word faster.

Result 2: Having empty speaker turns slows people down: compare clean vs. maxclean.

This is probably why the no-fillers, no-fillers/edits conditions are slow also.

In future experiments, we will omit empty speaker turns.

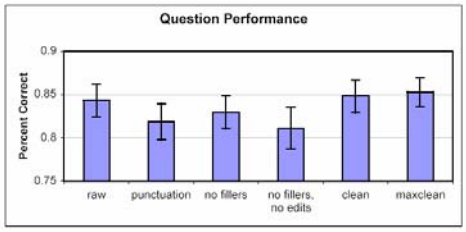
Experiment 4: Self-paced, 3 words at a time

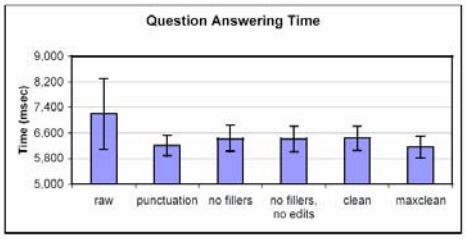


Result: Cleaned-up texts are read 24% faster on average, without loss in response-accuracy to questions. (105 seconds / text down to 80 seconds / text)

Note: These data seem to pattern like the rating data, but there is little correlation item-by-item between rating and RT. Therefore, faster RTs are not likely to be the cause of better ratings.

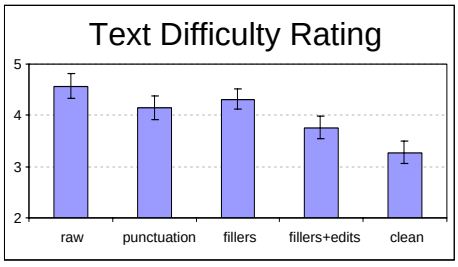
Experiment 4: Self-paced, 3 words at a time

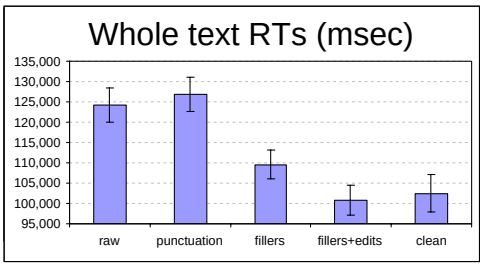




Null effects in response accuracies to questions and question response times: ceiling effects

Experiment 5: Self-paced, 1 word at a time



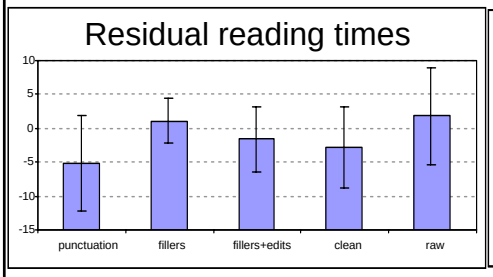


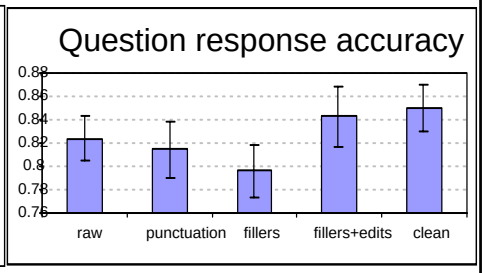
Same pattern of results in difficulty ratings and overall text RTs in Experiment 5 as in Experiment 4.

Cleaned-up texts are read 19% faster on average, without loss in response-accuracy to questions. (125 seconds / text down to 102 seconds / text)

Note: The texts are read more slowly overall when read one-word-at-a-time compared with being read three-words-at-a-time

Experiment 5: Self-paced, 1 word at a time





Null effects in RTs / word, response accuracies to questions, and question response times.

Potential Experiments

- **"Speed Test"**
  - **Subjects read as much as possible in a fixed amount of time**
  - **Eg.: $10 base compensation, $1 for each correct answer, -$0.25 for each incorrect answer**
- **Integrated workflow test**
  - **Insert raw, cleaned up transcripts into a simulated workflow environment TBD**
  - **Measure impact of transcript condition in workflow terms**

Work in Progress

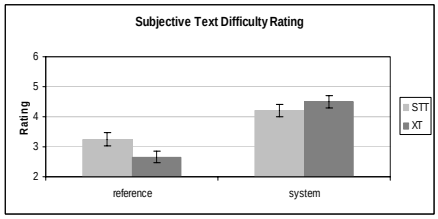
- Analyze data from moving-window experiments region by region, to see if some kinds of clean-up speed up or slow down reading times
- Moving window experiment in which speaker change information is a factor
- Writing a manual for the MIT mark-up conventions
- Evaluating the cross-coder reliability of the MIT mark-up conventions
- Mapping MIT mark-up to SimpleMDE5

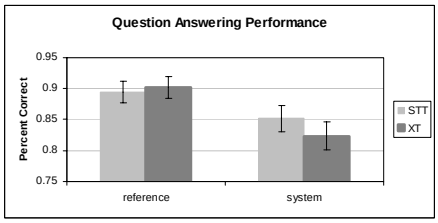
Appendix

- Results from Phase 1 experiments

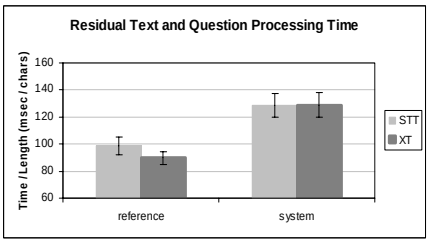
Form I Text Presentation

Text and Questions Together



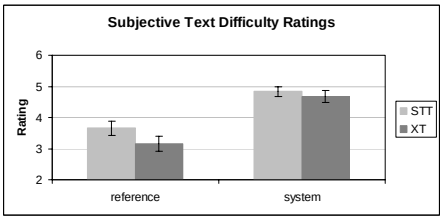


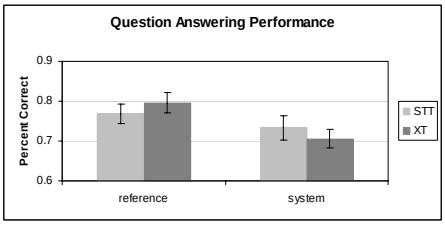
1. System Error has major impact on readability
2. Significant difference in subjective text difficulty rating in reference texts
3. Numerical (but not significant) difference in favor of cleaned up text for reference texts for question answering and text processing time
4. No clean-up differences in question answering time
5. System XT – numerical trend in opposite direction (errorful cleanup hurts readability)



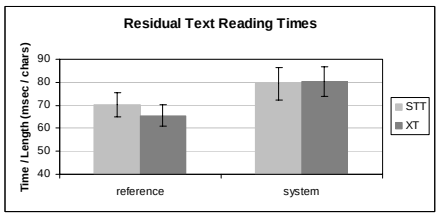
Form II Text Presentation

Text and Questions Separate





Similar to Form I, but stronger numerical trend favoring XT for question answering performance in the reference texts

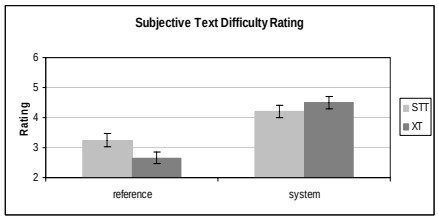


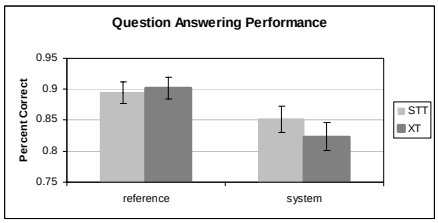
Introduction

- Primary uses of adding MDE to STT
  - Allow for rendering of "cleaned up" transcripts for human readers
  - Provide extra information for down-stream automated processes
- XT = "Cleaned up" STT Transcripts
  - Disfluencies removed
  - Punctuation and capitalization added for SUs
  - Maximally fluent, given the input
- Hypotheses:
  - XT is more readable for humans
  - Reference texts are more readable than system output texts

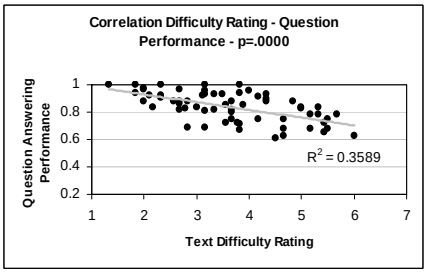
Form I Text Presentation

Text and Questions Together



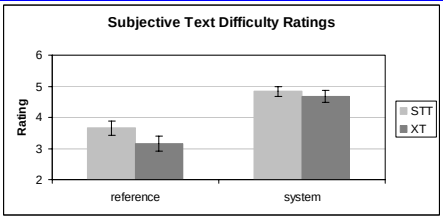


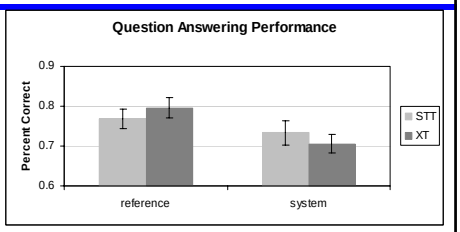
1. System Error has major impact on readability
2. Significant difference in subjective text difficulty rating in reference texts
3. Numerical (but not significant) difference in favor of cleaned up text for reference texts for question answering
4. Significant correlation between text difficulty and question answering performance
5. Task difficulty (answering the questions) could be influencing text difficulty ratings
6. No clean-up differences for question answering (possible ceiling effect)



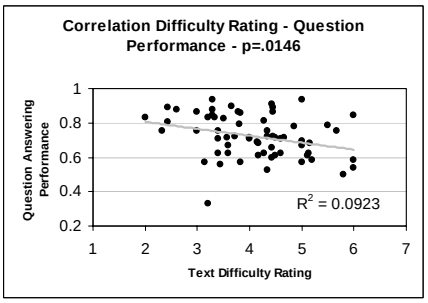
Form II Text Presentation

Text and Questions Separate





1. **Similar to Form I, but…**
2. **Measure of text difficulty instead of task difficulty**
3. **Stronger numerical trend favoring XT for question answering performance**
4. **System XT – numerical trend in opposite direction (errorful cleanup hurts readability)**
5. **Word error rate affects readability, in terms of difficulty ratings and text reading times (see appendix for correlation graphs)**
6. **Correlation of text difficulty rating with question answering performance suggests that text difficulty ratings are indicative of text readability (not just a subjective measure of how subjects like the texts)**



END OF MAIN SLIDES

BACKUP SLIDES FOLLOW

Proposed work:

Use more natural, more highly disfluent transcripts

- **The switchboard conversations that we have been analyzing are potentially not ecologically valid, because of the method in which they are gathered: two people forced to talk about some topic.**
- **Same methods applied to more naturalistic conversations. E.g., from meetings.**

Outline for Readability Update

- Definitions and Design
  - Measuring readability
  - Conceptual framework for STT enhancement
  - Speaker turn / "blob of words" experiment
  - Improved experimental design
  - Passage-level vs word-level forms for test trials

  Eurospeech 2003 paper
- Current Results
  - Word-level forms for test trials (three word window)
  - Cleanup improves readability for reference transcripts
  - Simplified metadata annotation was used
  - Work in progress: one-word window (for Nov PI meeting)
- Next Steps / Proposals
  - Word-level speaker turn experiment (Jan PI meeting)
  - Readability of RT-03F system output (May PI meeting)
  - Machine Translation experiments? (June MT-04 workshop)
- Appendix

Backup Slides with Transcript Examples

- *We ended up using these, but during the break immediately after our talk.*
- *We skipped over the introductory "Comments on 'cleaned up' transcripts" slides, going directly to the four-way comparison of the "cow manure" transcript in several conditions, the MIT version being arguably the most readable.*

Comments on "cleaned up" transcripts

- **Results reported at May PI Meeting and at Eurospeech'03 did not show objective differences in verbatim vs. "cleaned up" transcripts.**
- **Goal: give the experiments a better chance to measure differences in "maximally readable" transcripts.**
- **Re-examined experimental design:**
  - **Improvements to experimental protocol with self-paced reading.**
- **Re-examined experimental data:**
  - **Very careful examination of the 18 x ~250 word transcript passages drawn from Shriberg's selection of 40 SWB/CTS transcripts.**
  - **Estimation of final transduction step from "Rich Transcript" (RT) to "Transformed Transcript" (XT) directly from verbatim transcripts.**
  - **Very closely related to LDC SimpleMDE v5 spec but even more simplified…**
  - **But also includes more extensive standardized orthography, including commas.**
- **Next steps:**
  - **Establishing relationship to EARS MDE tasks and transduction from SimpleMDE v5.**
  - **Experiments with system output will use closest transduction from SimpleMDE v5.**
  - **Experiments with system output will include control cases using closest transduction from reference transcripts from SimpleMDE v5.**
- **Comments:**
  - **Creating fully "cleaned-up" human readable output is not a currently defined task for EARS/RT.**
  - **Allowed for changes to SimpleMDE v5 -> v6, v7, …**
  - **Consider relationship to a new "cleaned up" transcript task for RT-05.**

Sample Transcript for Cleanup Transduction

- **Sample CTS conversation from RT03F training data: sw2648**
- **Demonstration that additional work is needed for transduction step. Four options:**
  1. **Unmodified Rendering from MDETool (too cluttered and long)**
  2. **Direct Transduction, keeping all backchannels (too cluttered)**
  3. **Direct Transduction, dropping all backchannels (too dense)**
  4. **MIT Transduction, keeps <u>some</u> backchannels (easier to read)**

Annotation Rendering in MDETool (1)

A: it's in Dublin Texas /.
A: i'm not really sure what the county it's Stephenville /.
B: uh-huh /@
B: oh /@
B: okay /@
B: okay /@
A: okay /@
A: where Stephenville's at they've had a lot of problems /.
A: because they've introduced a lot of dairies there /.
B: uh-huh /@
B: yeah /@
A: so they have a lot of cattle in the area /.
A: per lot they've got like a lot of head of cattle /.
A: because_1 it's not like it was a big ranch where they let the cow roam around free /.
A: they've got to have them there to be able to milk them /.
B: right /@
B: uh-huh /@
A: and i read in this article where one cow a day produces like a hundred or so pounds of manure /.
B: oh /@
A: one cow /.
A: i was like /,
A: my God /.
A: and this thing says /,
A: that people think /,
A: cow manure is good for fertilizer /.
A: but when you get that much manure /,
A: it says /,
A: it becomes a real problem /.
A: because it's not /.
A: when you buy cow manure at the store /,
A: they've added stuff to it /&
A: and they've added humus and stuff that breaks it down /.
B: uh-huh /@

Annotation Rendering in MDETool (2)

A: and this article said /,
A: that they've got a real problem /&
A: that you've got toxins and bacteria in the manure /.
A: what it's doing is /,
A: it's going through they're water /,
B: uh-huh /@
A: they're natural water /.
A: i forget what that's called /.
A: it's like a spring water /&
A: but it's below the surface /.
A: it's not very far down /.
B: yeah /@
B: uh-huh /@
A: and all this stuff is seeping through to the water /.
B: oh /@
A: and they're afraid that within a certain amount of years that they're water in that town will be totally polluted /&
A: and they won't be able to have any drinking water /.
A: because they will have polluted /,
B: it will /?
A: completely polluted so that there's too many toxins and bacterial growth /.
B: uh-huh /@

Direct Transduction with Backchannels

A: It's in Dublin Texas. I'm not really sure what the county it's Stephenville.
B: Oh. Okay.
A: Okay.
B: Okay.
A: Where Stephenville's at they've had a lot of problems.
B: Uh-huh.
A: Because they've introduced a lot of dairies there.
B: Yeah.
A: So they have a lot of cattle in the area. per lot they've got like a lot of head of cattle. Because it's not like it was a big ranch where they let the cow roam around free.
B: Right.
B: Uh-huh.
A: They've got to have them there to be able to milk them. and i read in this article i couldn't believe it where one cow a day produces like a hundred or so pounds of manure.
B: Oh.
A: One cow. I Was like my God. and this thing says that people think cow manure is good for fertilizer. But when you get that much manure it says it becomes a real problem. Because it's not. when you buy cow manure at the store they've added stuff to it and they've added humus and stuff that breaks it down.
B: Uh-huh.
A: and this article said that they've got a real problem that you've got toxins and bacteria in the manure. what it's doing is it's going through they're water
B: Uh-huh.
A: they're natural water. i forget what that's called. it's like a spring water but it's below the surface.
B: Yeah.
A: It's not very far down.
B: Uh-huh.
A: And all this stuff is seeping through to the water.
B: Oh.
A: And they're afraid that within a certain amount of years that they're water in that town will be totally polluted and they won't be able to have any drinking water.
B: It will ?
A: Because they will have polluted completely polluted so that there's too many toxins and bacterial growth.
B: Uh-huh.

Direct Transduction, no Backchannels

A: It's in Dublin Texas. I'm not really sure what the county it's Stephenville. Where Stephenville's at they've had a lot of problems. Because they've introduced a lot of dairies there. So they have a lot of cattle in the area. per lot they've got like a lot of head of cattle. Because it's not like it was a big ranch where they let the cow roam around free. They've got to have them there to be able to milk them. and i read in this article i couldn't believe it where one cow a day produces like a hundred or so pounds of manure. One cow. I Was like my God. and this thing says that people think cow manure is good for fertilizer. But when you get that much manure it says it becomes a real problem. Because it's not. when you buy cow manure at the store they've added stuff to it and they've added humus and stuff that breaks it down. and this article said that they've got a real problem that you've got toxins and bacteria in the manure. what it's doing is it's going through they're water they're natural water. i forget what that's called. it's like a spring water but it's below the surface. It's not very far down. And all this stuff is seeping through to the water. And they're afraid that within a certain amount of years that they're water in that town will be totally polluted and they won't be able to have any drinking water. Because they will have polluted completely polluted so that there's too many toxins and bacterial growth.

B: It will ?

MIT Transduction

A: It's in Dublin, Texas. I'm not really sure what the county, it's Stephenville.

B: Okay.

A: Okay, where Stephenville's at they've had a lot of problems because they've introduced a lot of dairies there. So they have a lot of cattle in the area. They've got a lot of head of cattle because, it's not like it was a big ranch where they let the cow roam around free.

B: Right.

A: They've got to have them there to be able to milk them, and I read in this article, I couldn't believe it, one cow a day produces a hundred or so pounds of manure. One cow, I was like, my god, and this thing says that people think cow manure is good for fertilizer, but when you get that much manure, it says it becomes a real problem because when you buy cow manure at the store they've added stuff to it, and they've added humus and stuff that... breaks it down, and this article said that they've got a real problem, that you've got toxins and bacteria in the manure. That is what it's doing is it's going through, they don't have a very deep ... their natural water, I forget what that's called, it's like spring water. But it's below the surface. It's not very far down. And all this stuff is seeping through to the water.

A: And they're afraid that within a certain amount of years, their water in that town will be totally polluted.

B: It will.

A: And they won't be able to have any drinking water because they will have completely polluted so that there's too many toxins and bacterial growth.

Comparison for MIT cleanup transduction

- How different is the MIT transduction from some possible direct transductions?
- Using WER (sclite) as a measure of closeness:

Ref -> Hyp:	MIT	Direct, -bc	Direct, +bc
MIT	0%	23%	28%
Direct, -bc	22%	0%	17%
Direct, +bc	31%	20%	0%

Comments on “cleaned up” transcripts

- Next steps:
  - Establishing relationship to EARS MDE tasks and transduction from SimpleMDE v5.
  - Experiments with system output will use closest transduction from SimpleMDE v5.
  - Experiments with system output will include control cases using closest transduction from reference transcripts from SimpleMDE v5.
- Review main comments:
  - Learning how to create maximally readable transcripts from EARS RT markup may be a research task in itself
  - Consider relationship to a new “cleaned up” transcript task for RT-05.